

Universidade Federal do Rio Grande do Sul
Programa de Pós Graduação em Computação
Instituto de Informática

Earth Simulator

por: Augusto Ordobás Bortolás
prof. Philippe Olivier Alexandre Navaux

maio/2005

1. Introdução

O Earth Simulator (ES), supercomputador vetorial altamente paralelo de memória distribuída, foi construído em 2002 pela NEC para o Earth Simulator Center, em Yokohama, com o objetivo de efetuar simulações e previsões de clima e terremotos. A intrincada relação entre a atmosfera terrestre, os oceanos e a crosta terrestre é impossível de ser recriada em laboratório. A única forma capaz de dar ao homem um melhor entendimento dessas relações é através da simulação, que verifica como a alteração em uma ou mais variáveis influencia todas as demais. Mas mesmo os supercomputadores até agora existentes eram incapazes de efetuar tamanha quantidade de cálculos.

O Earth Simulator trabalha através da criação de uma teia virtual, partindo do Equador e se estendendo sobre a toda a superfície da Terra. Seu poder de cálculo é capaz de analisar os movimentos das correntes marítimas num período de mil anos e efetuar previsões de tempo e de terremotos de até 180 dias.

O restante do texto está organizado da seguinte maneira: a seção 2 expõe as características de *hardware* do ES, bem como sua arquitetura e organização; a seguir, na seção 3, são expostas as configurações de *software* do ES, como sistema operacional; resultados de performance são ilustrados na seção 4; finalmente, o trabalho é concluído na seção final (5), e são apresentadas as referências bibliográficas (seção 6).

2. Hardware

2.1. Configuração do Sistema

O Earth simulator é um supercomputador vetorial altamente paralelo de memória distribuída, consistindo em 640 nós de processamento (Pns), conectados por 640x640 *switches* de barra cruzada de um estágio. Cada nó possui 8 processadores vetoriais aritméticos, 16GB de memória compartilhada, uma unidade de controle de acesso remoto (RCU) e um processador de I/O. As figuras 1 e 2 ilustram como estão interligados os nós à rede de comunicação. Como está exposto nestas figuras, a rede conecta os nós diretamente, provendo conexões de alta velocidade.

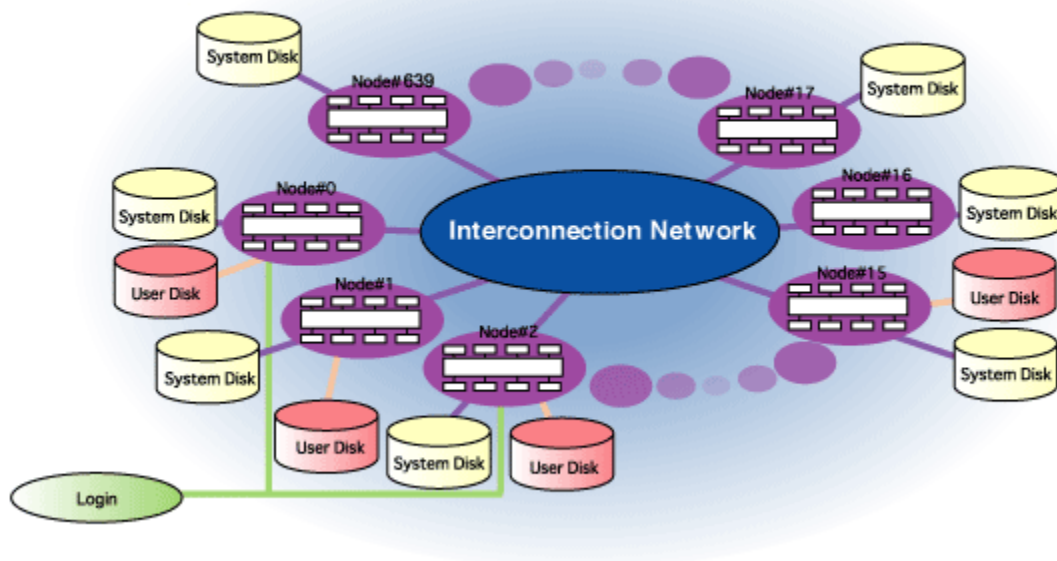


Figura 1 – Interconexão dos nós

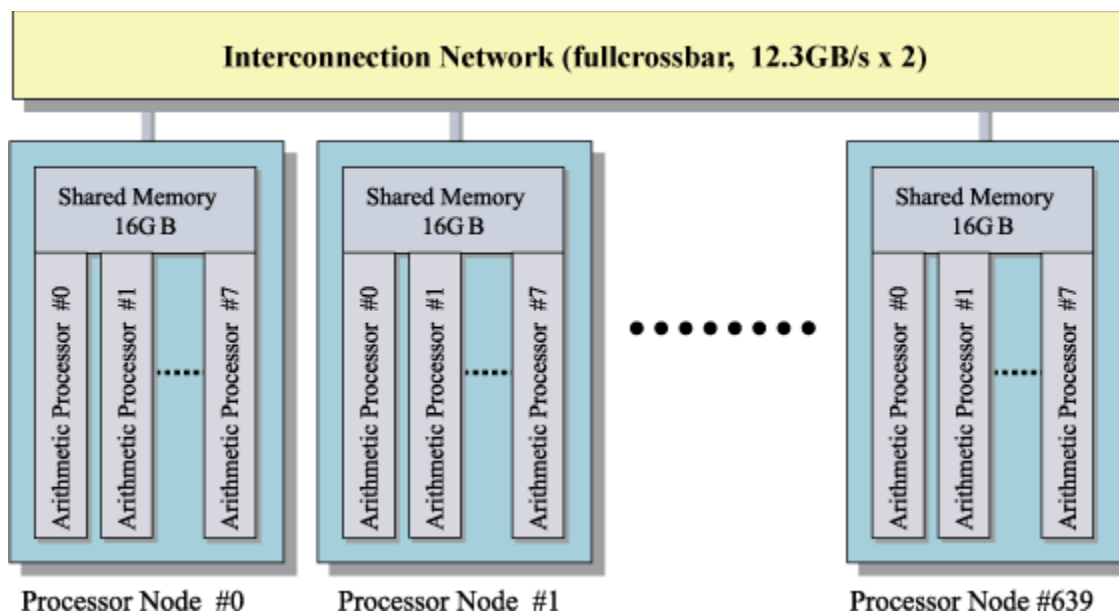


Figura 2 – Estrutura dos nós e suas interconexões

2.2. Processador Vetorial Aritmético

Cada processador vetorial aritmético (AP) consiste em uma unidade super escalar (SU), uma unidade vetorial (VU) e uma unidade de controle de acesso à memória principal em um único chip. O AP opera a uma frequência de 500MHz com alguns circuitos operando a 1GHz. Cada SU é um processador super escalar com caches de instruções de 64KB, caches de dados de 64KB e 128 registradores escalares de propósito geral. O processador também possui predição de desvio e pré-busca de dados.

Cada VU possui 72 registradores vetoriais, cada um com 256 posições, e possui 8 conjuntos de 6 diferentes tipos de pipelines vetoriais: adição/shifting, multiplicação, divisão, operações lógicas, masking, load/store. Os mesmos tipos de pipelines vetoriais trabalham em conjunto em um vetor de instruções, enquanto os diferentes tipos de pipelines podem operar concorrentemente. A figura 3 ilustra a organização lógica do AP. Na figura 4 é exposta a organização física do AP.

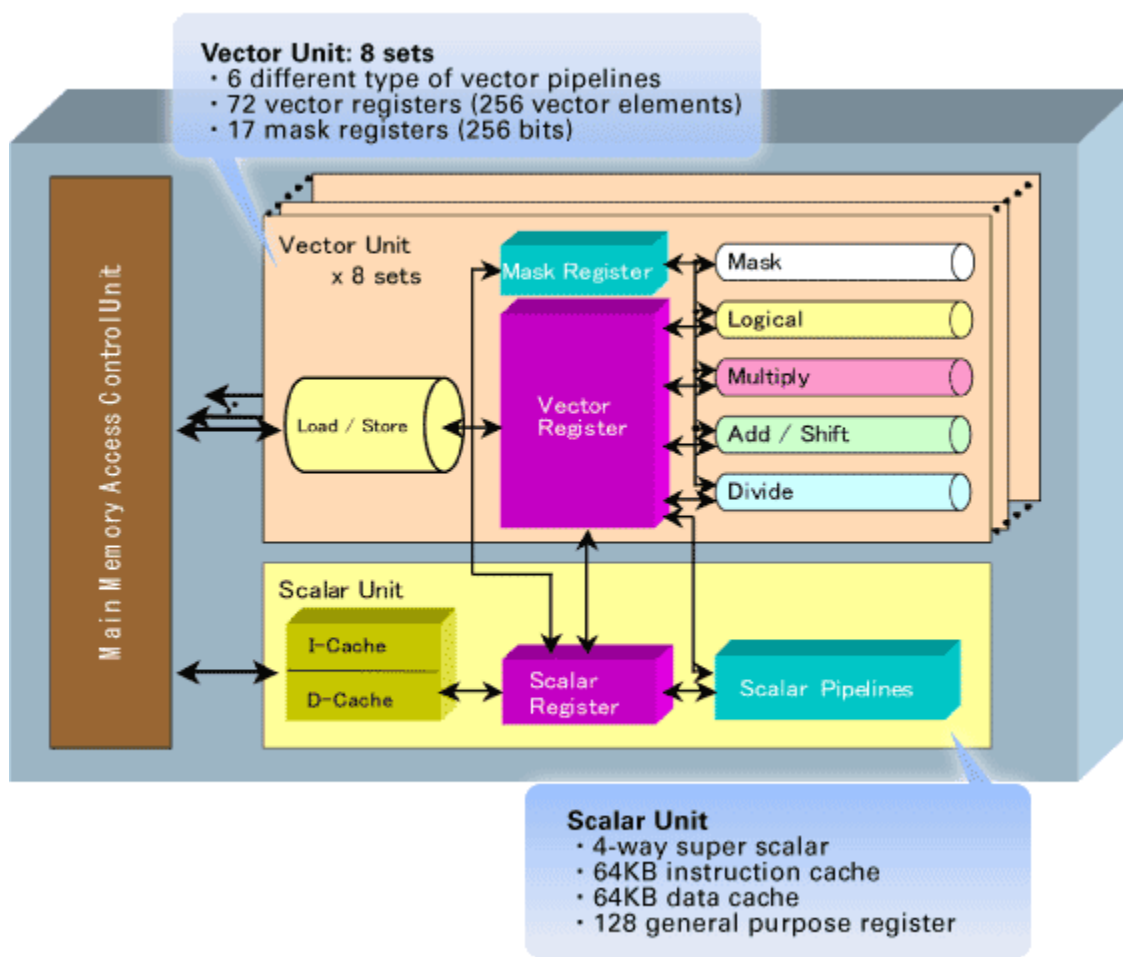


Figura 3 – Organização lógica do AP

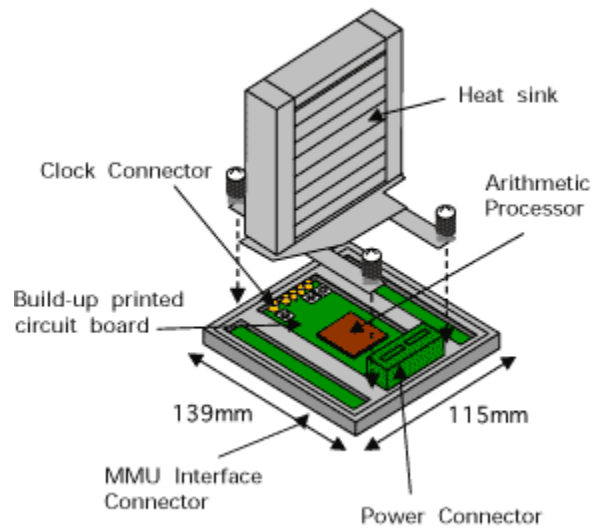


Figura 4 – Organização física do AP

2.3. Processor Node (Nó)

A memória principal de cada nó é dividida em 32 blocos, como ilustra a figura 5. Cada AP possui conexão física direta a estes blocos.

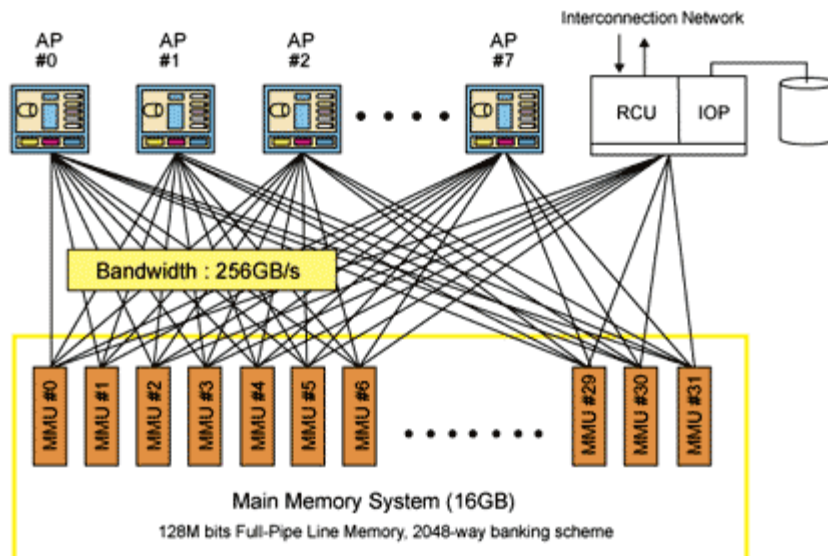


Figura 5 – Memória principal do Nó

A ideia central da fig. 5 é ilustrar a organização da memória principal dentro de um nó. Também é mostrado como é feito o acesso remoto aos dados em memórias distribuídas, o qual é feito através da unidade de controle remoto (RCU).

2.4. Rede de Interconexão

A RCU está diretamente conectada a chaveadores de barra cruzada e controla os dados de comunicação entre os nós a uma taxa de 12,3GB/s, em transferências bidirecionais. A largura total da banda da rede entre os nós chega a 8TB/s.

Muitos modos de transferências de dados, incluindo acesso a sub-arrays tridimensionais e modos de acesso indiretos são realizados via *hardware*. Em uma operação que envolve acesso a um dado de um sub-array, o dado é movido de um nó a outro em uma única operação de *hardware*, e relativamente pouco tempo é consumido nesse processamento.

Cada cabine de rede (são 64 ao total), são formadas por duas unidades de chaveadores de dados de 640X640. Existe uma cabine com duas unidades de controle de rede, como ilustram as figuras 6 e 7.

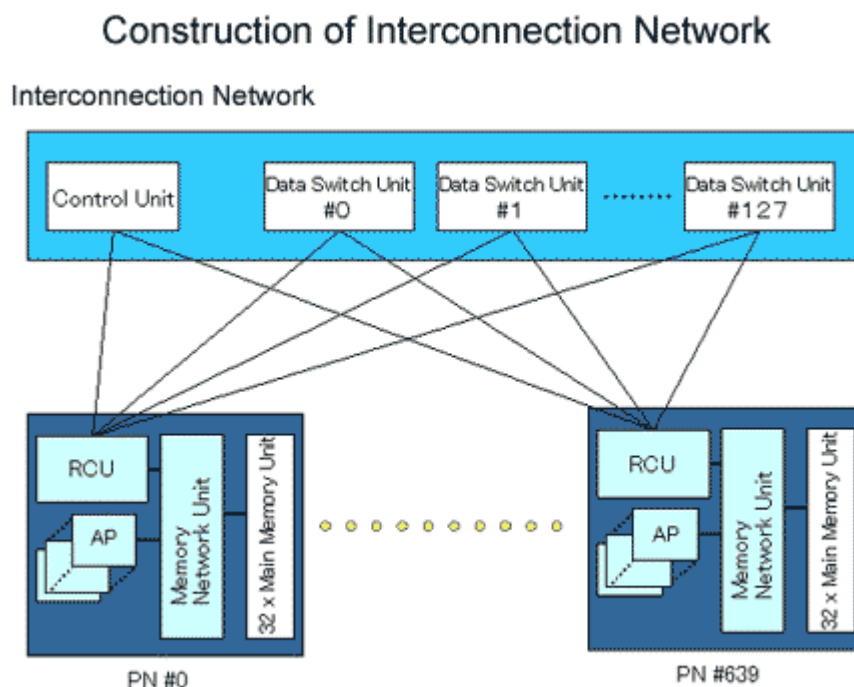


Figura 6 – Estrutura da rede de interconexão

Como ilustra a figura 7, existem portanto 130 unidades de rede (65 cabines), interligadas com as 640 cabines de nós, totalizando a incrível quantidade de 83.200 cabos elétricos.

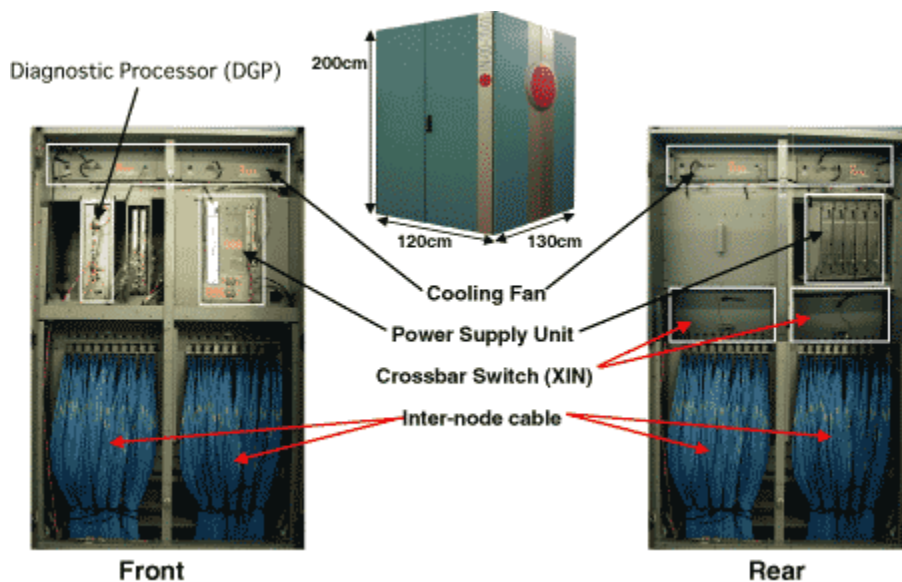


Figura 9 – Estrutura física da cabine de rede

2.5. Sistema de Processamento Massivo de Dados (MDPS)

O MDPS (Mass Data Processing System), instalado em outubro de 2003, é um sistema de processamento massivo de dados, o qual consiste em 4 processadores de serviço de arquivos (FSPs), 250TB de disco rígido e 1.5PB de bibliotecas de fitas de cartucho (CTL). O objetivo do MDPS foi aumentar a vazão da transferência de dados e a acessibilidade aos dados.

É interessante destacar os seguintes itens:

- A velocidade de transferência de salvamento e extração de dados entre o ES e o sistema de armazenamento em massa de dados se torna de 2 a 5 vezes mais rápido. Essa melhoria foi realizada através da expansão das habilidades dos cabos de transferência e substituição do arquivo de fitas por discos.
- A visão do usuário dos sistema de armazenamento são discos rígidos enormes, os quais podem ser manipulados com comandos e ferramentas usuais do UNIX, através do servidor de *login*.
- O MDPS torna os acessos de usuários tão demorados quanto os resultados providos pelo ES, pois pode transferir dados a servidores dedicados fora da LAN do ES, como mostra a figura 10.

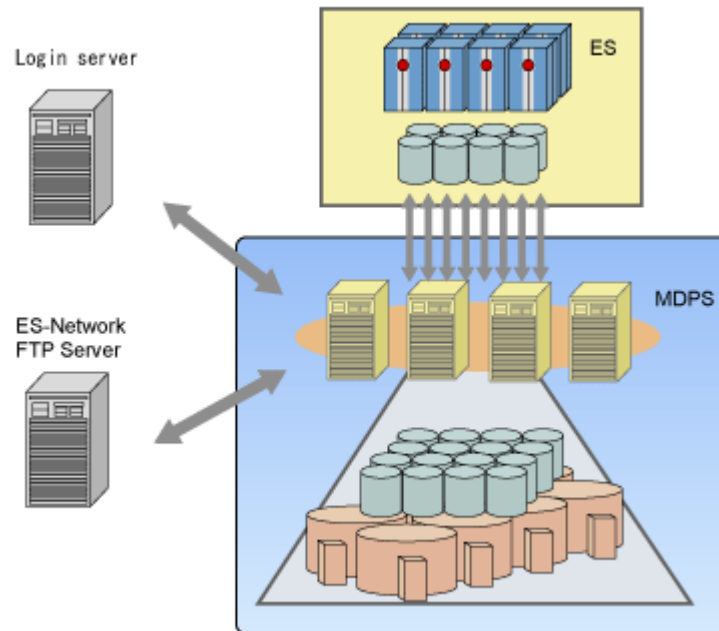


Figura 10 – Acesso dos usuários ao MDPS

A figura 11 ilustra as várias redes ao qual a rede do ES possui acesso.

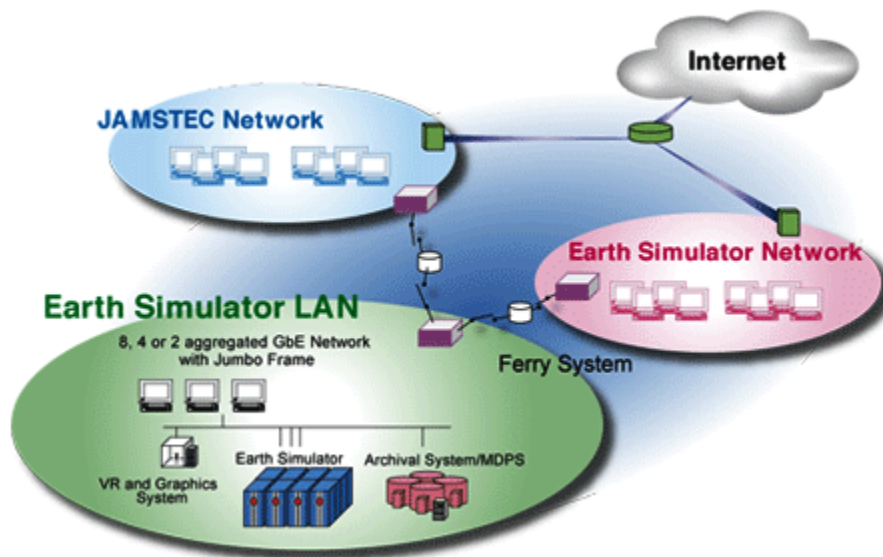


Figura 11 – Redes acessadas pelo ES

3. Software

Todos os componentes de *software* do ES foram desenvolvidos especialmente para o mesmo, deste modo toda a sua potencialidade pode ser explorada em sua totalidade. Nas subseções seguintes são expostas as descrições dos *softwares* usados no SO, sistema de arquivos paralelo, escalonador de tarefas e ambiente de programação.

3.1. Sistema Operacional

O sistema operacional usado no ES é uma versão melhorada do SUPER-UX, sistema operacional baseado em UNIX da NEC, desenvolvidos para a série de supercomputadores SX.

A principal mudança feita no SUPER-UX foi prover escalabilidade, para os 640 nós do ES. Também foram otimizados os gerenciamentos de processos, memória e arquivos.

O SUPER-UX já provê algumas características que são extremamente importantes ao ES, como comunicação veloz entre os nodos através da rede, espaço de endereçamento global através dos nós e um sistema de *cluster*, como ilustra a figura 12.

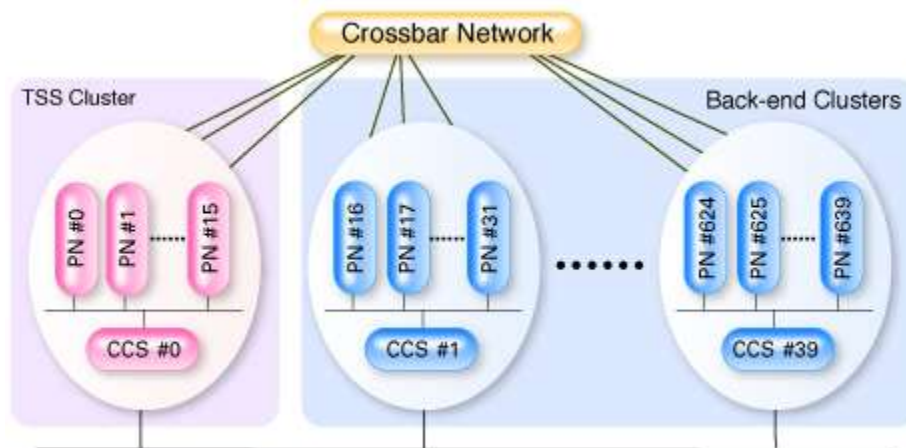


Figura 12 – Sistema de *clusters* do ES

Um sistema de gerenciamento hierárquico foi introduzido no ES, para prover um melhor aproveitamento do seu alto poder de processamento. Cada 16 nós formam um *cluster*, formando 40 *clusters* no total. Quando uma tarefa relativamente pequena é escalonada, apenas alguns *clusters* são alocados, oferecendo assim a possibilidade de reaproveitar os nós que estão sem trabalho.

Cada *cluster* possui uma estação de controle de *cluster* (CSS) o qual monitora o

estado dos nós. Existe uma super estação de controle de *cluster* (SCCS), que desempenha tarefa importante na integração e coordenação de todas as operações dos CSSs.

3.2. Sistema de Arquivos Paralelo

O sistema de arquivos paralelo do ES (PFS) provê características de I/O paralelas ao ES, como gerenciamento de múltiplos arquivos, os quais estão alocados em diferentes discos de diferentes nós, e são logicamente um único arquivo. O PFS é ilustrado na figura 13.

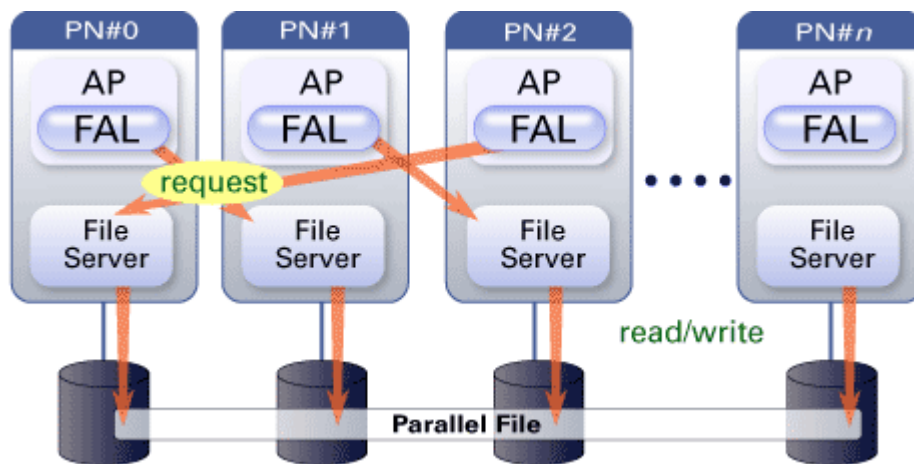


Figura 12 – Sistema de arquivos paralelo

Cada processo de um programa paralelo pode ler/escrever em dados distribuídos de/para o arquivo paralelo concorrentemente com alta performance e usabilidade de I/O. Como mostrado na figura 12, um arquivo pode ser quebrado em pedaços e armazenado paralelamente em vários nós do sistema. Quando um programa acessa o arquivo, a biblioteca de acesso a arquivos (FAL) envia uma requisição de I/O pela rede para o servidor de arquivos do nó que possui o trecho de dados do arquivo a ser acessado.

3.3. Modelo de Programação no ES

Para se obter o melhor desempenho que o ES pode oferecer, duas características importantes devem ser exploradas: memória compartilhada dentro dos nós e memória distribuída entre os nós. Existem compiladores Fortran 90, C e C++ para o ES. Podem ser usadas as bibliotecas MPI (Message Passing Interface) e HPF (High Performance

Fortran) para programação paralela.

Também estão disponíveis várias ferramentas para desenvolvimento, depuração e ajuste de programas: Batch Debug (Debugging), PSUITE (development environment), Vampir (performance analysis tool), HPFPROF (performance analysis tool).

4. Performance

A performance máxima, teoricamente, do ES é 40TFlops no total, 8GFlops por cada AP e 64GFlops por nó. O ES possui 5120 APs no total. Possui uma memória de 10TB no total.

No entanto, a performance máxima alcançada até hoje, usando o *benchmark* Linpack, foi de 35.86TFlops. Uma eficiência de 87.5%.

A figura 13 ilustra testes de desempenho realizados com o ES.

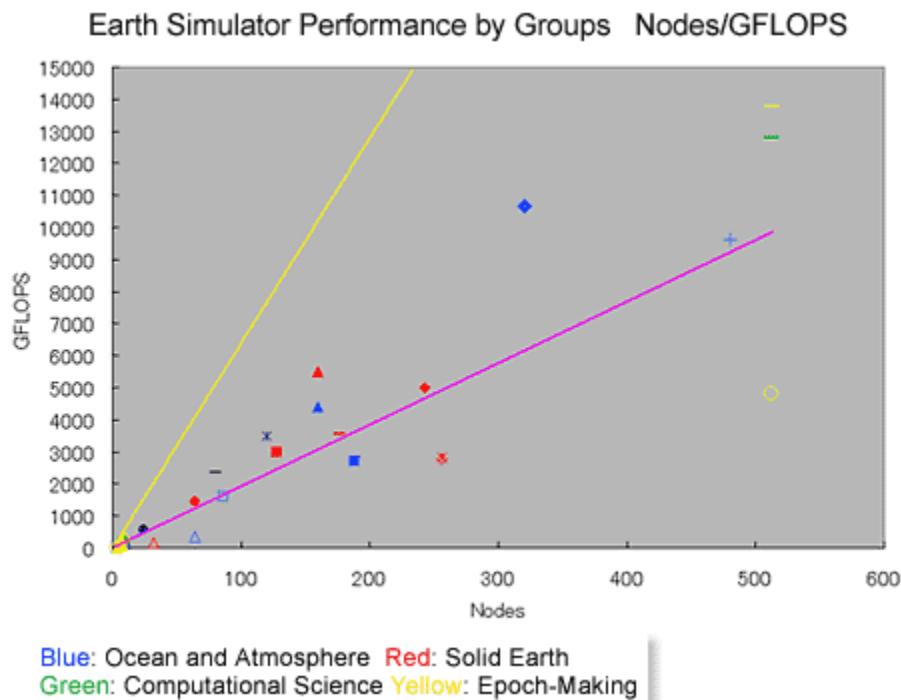


Figura 13 – Testes de Desempenho no ES

A linha amarela indica o máximo de desempenho na teoria. A linha rosa indica 30% da performance alcançada.

5. Conclusões

Os supercomputadores japoneses ficaram famosos por usarem tecnologia vetorial, e o ES foi um deles. Para se ter uma idéia do potencial do ES, em 2003, quando foi lançado, alcançou uma performance 5 vezes maior que o supercomputador mais potente da época, o qual estava no topo da lista da top500.

O ES demonstrou todo o poder que o paralelismo pode alcançar, e que várias características arquiteturais podem ser combinadas para se alcançar alto processamento:

- Processamento vetorial;
- Memória distribuída entre nós;
- Memória compartilhada entre processadores em um nó;
- Redes de conexão de alta performance.

Mesmo tendo sido desbancado por dois computadores norte-americanos, até o momento, o ES ainda é visto como uma obra de respeito dos japoneses.

6. Referências

Página oficial do Earth Simulator:

<http://www.es.jamstec.go.jp/esc/eng/index.html>

TOP500:

<http://www.top500.org>

Site de notícias sobre inovações tecnológicas:

<http://inovacaotecnologica.com.br>